

# Introduction to spatial data and models

Andrew O. Finley<sup>1</sup> and Alan Gelfand<sup>2</sup>

<sup>1</sup> Department of Forestry, Michigan State University, Lansing, Michigan.

<sup>2</sup> Department of Statistical Science, Duke University, Durham, North Carolina.

September 9, 2014

1

- Researchers in diverse areas such as ecology, forestry, climatology, and environmental health, are increasingly faced with the task of analyzing data that are:
    - highly multivariate, with many important predictors and response variables,
    - geographically referenced, and often presented as maps, and
    - temporally correlated, as in longitudinal or other time series structures.
- ⇒ motivates **hierarchical** modeling and data analysis for complex spatial (and spatiotemporal) data sets.

2

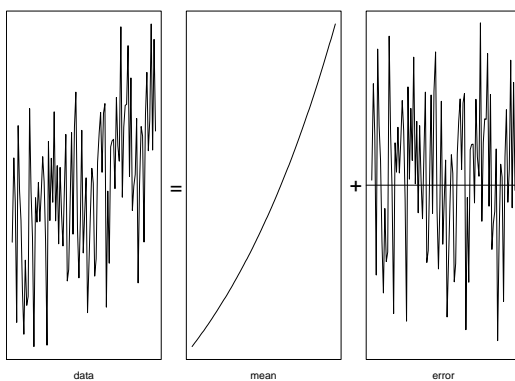
- **point-referenced data**, where  $y(\mathbf{s})$  is a random vector at a location  $\mathbf{s} \in \mathbb{R}^r$ , where  $\mathbf{s}$  varies **continuously** over  $D$ , a fixed subset of  $\mathbb{R}^r$  that contains an  $r$ -dimensional rectangle of positive volume;
- **areal data**, where  $D$  is again a fixed subset (of regular or irregular shape), but now partitioned into a **finite** number of areal units with well-defined boundaries;
- **point pattern data**, where now  $D$  is itself random; its index set gives the locations of random events that are the spatial point pattern.  $y(\mathbf{s})$  itself can simply equal 1 for all  $\mathbf{s} \in D$  (indicating occurrence of the event), or possibly give some additional covariate information (producing a **marked point pattern process**).

3

- First step in analyzing data
- First Law of Geography: Mean + Error
- Mean: first-order behavior
- Error: second-order behavior (covariance function)
- EDA tools examine both first and second order behavior
- Preliminary displays: Simple locations to surface displays

4

## First Law of Geography



5

## Scallops Sites

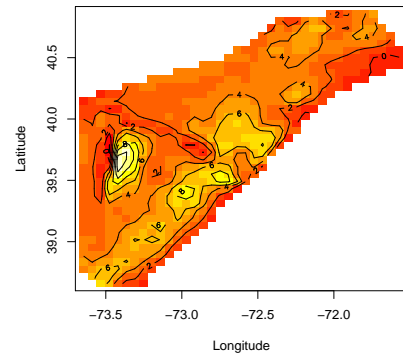


6

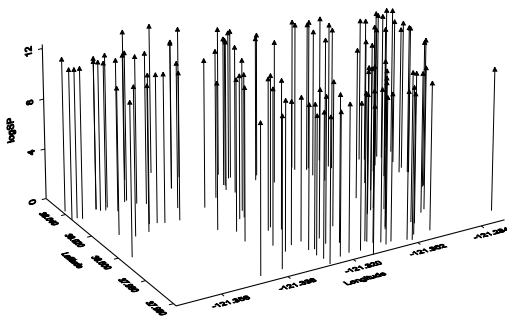
- Spatial surface observed at finite set of locations  $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$
- Tessellate the spatial domain (usually with data locations as vertices)
- Fit an interpolating polynomial:

$$f(\mathbf{s}) = \sum_i w_i(\mathcal{S}; \mathbf{s}) f(\mathbf{s}_i)$$

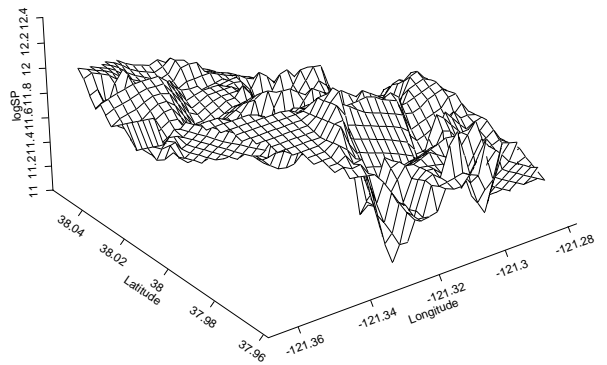
- "Interpolate" by reading off  $f(\mathbf{s}_0)$ .
- Issues:
  - Sensitivity to tessellations
  - Choices of multivariate interpolators
  - Numerical error analysis



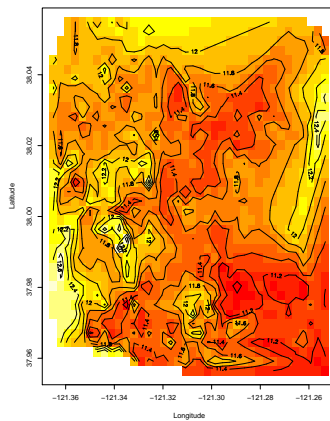
### Drop-line scatter plot



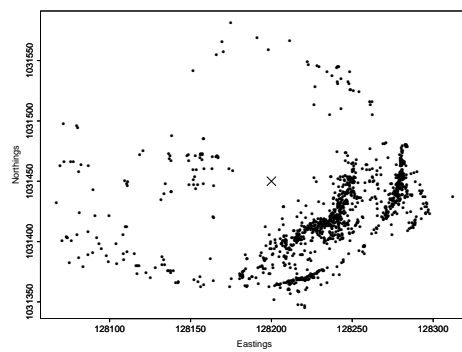
### Surface plot



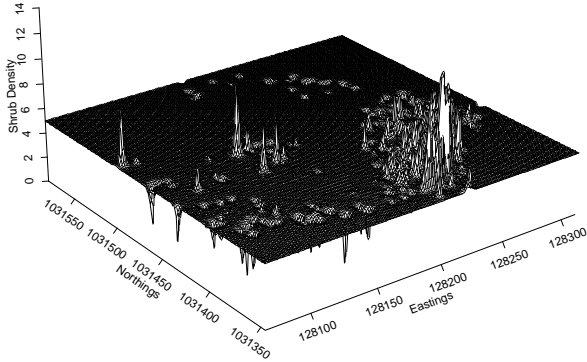
### Image contour plot



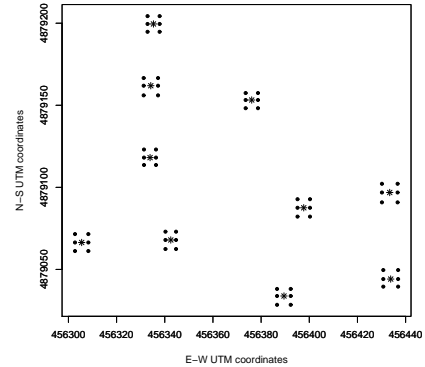
### Locations form patterns



### Surface features



### Interesting plot arrangements



- Point-level modeling refers to modeling of spatial data collected at locations referenced by **coordinates** (e.g., lat-long, Easting-Northing).
- **Fundamental concept:** Data from a spatial process  $\{y(\mathbf{s}) : \mathbf{s} \in D\}$ , where  $D$  is a fixed subset in Euclidean space.
- **Example:**  $y(\mathbf{s})$  is a **pollutant level** at site  $\mathbf{s}$
- **Conceptually:** Pollutant level exists at all possible sites
- **Practically:** Data will be a partial realization of a spatial process – observed at  $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$
- **Statistical objectives:** **Inference** about the process  $y(\mathbf{s})$ ; **predict** at new locations.

Suppose our spatial process has a mean,  $\mu(\mathbf{s}) = E(y(\mathbf{s}))$ , and that the variance of  $y(\mathbf{s})$  exists for all  $\mathbf{s} \in D$ .

- **Strong stationarity:** If for any given set of sites, and any displacement  $\mathbf{h}$ , the distribution of  $(y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))$  is the same as,  $(y(\mathbf{s}_1 + \mathbf{h}), \dots, y(\mathbf{s}_n + \mathbf{h}))$ .
- **Weak stationarity:** Constant mean  $\mu(\mathbf{s}) = \mu$ , and  $Cov(y(\mathbf{s}), y(\mathbf{s} + \mathbf{h})) = C(\mathbf{h})$ : the covariance depends only upon the displacement (or separation) vector.
- **Strong stationarity implies weak stationarity**
- The process is **Gaussian** if  $\mathbf{y} = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))$  has a **multivariate normal** distribution.
- **For Gaussian processes, strong and weak stationarity are equivalent.**

### Variograms

- Suppose we assume  $E[y(\mathbf{s} + \mathbf{h}) - y(\mathbf{s})] = 0$  and define

$$E[y(\mathbf{s} + \mathbf{h}) - y(\mathbf{s})]^2 = Var(y(\mathbf{s} + \mathbf{h}) - y(\mathbf{s})) = 2\gamma(\mathbf{h}).$$

This is sensible if the left hand side depends only upon  $\mathbf{h}$ . Then we say the process is **intrinsically stationary**.

- $\gamma(\mathbf{h})$  is called the **semivariogram** and  $2\gamma(\mathbf{h})$  is called the **variogram**.

Note that intrinsic stationarity defines **only** the first and second moments of the differences  $y(\mathbf{s} + \mathbf{h}) - y(\mathbf{s})$ . It says nothing about the **joint** distribution of a collection of variables  $y(\mathbf{s}_1), \dots, y(\mathbf{s}_n)$ , and thus provides **no likelihood**.

### Intrinsic Stationarity and Ergodicity

- Relationship between  $\gamma(\mathbf{h})$  and  $C(\mathbf{h})$ :

$$\begin{aligned} 2\gamma(\mathbf{h}) &= Var(y(\mathbf{s} + \mathbf{h})) + Var(y(\mathbf{s})) - 2Cov(y(\mathbf{s} + \mathbf{h}), y(\mathbf{s})) \\ &= C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h}) \\ &= 2[C(\mathbf{0}) - C(\mathbf{h})]. \end{aligned}$$

- Easy to recover  $\gamma$  from  $C$ . The converse needs the additional assumption of **ergodicity**:  $\lim_{\|\mathbf{u}\| \rightarrow \infty} C(\mathbf{u}) = 0$ .
- So  $\lim_{\|\mathbf{u}\| \rightarrow \infty} \gamma(\mathbf{u}) = C(\mathbf{0})$ , and we can recover  $C$  from  $\gamma$  as long as this limit exists.

$$C(\mathbf{h}) = \lim_{\|\mathbf{u}\| \rightarrow \infty} \gamma(\mathbf{u}) - \gamma(\mathbf{h}).$$

- When  $\gamma(\mathbf{h})$  or  $C(\mathbf{h})$  depends upon the separation vector only through the distance  $\|\mathbf{h}\|$ , we say that the process is **isotropic**. In that case, we write  $\gamma(\|\mathbf{h}\|)$  or  $C(\|\mathbf{h}\|)$ . Otherwise we say that the process is **anisotropic**.
- If the process is **intrinsically stationary** and **isotropic**, it is also called **homogeneous**.

Isotropic processes are popular because of their **simplicity**, **interpretability**, and because a number of relatively **simple parametric forms** are available as candidates for  $C$  (and  $\gamma$ ). Denoting  $\|\mathbf{h}\|$  by  $t$  for notational simplicity, the next two tables provide a few examples...

### Some common isotropic variograms

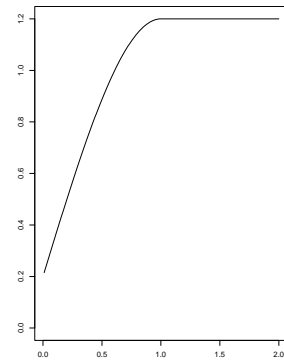
model	Variogram, $\gamma(t)$
Linear	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 t & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Spherical	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t \geq 1/\phi \\ \tau^2 + \sigma^2 \left[ \frac{3}{2}\phi t - \frac{1}{2}(\phi t)^3 \right] & \text{if } 0 < t \leq 1/\phi \\ 0 & \text{otherwise} \end{cases}$
Exponential	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi t)) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Powered exponential	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(- \phi t ^p)) & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$
Matérn at $\nu = 3/2$	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 [1 - (1 + \phi t) e^{-\phi t}] & \text{if } t > 0 \\ 0 & \text{o/w} \end{cases}$

### Examples: Spherical Variogram

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t \geq 1/\phi \\ \tau^2 + \sigma^2 \left[ \frac{3}{2}\phi t - \frac{1}{2}(\phi t)^3 \right] & \text{if } 0 < t \leq 1/\phi \\ 0 & \text{if } t = 0. \end{cases}$$

- While  $\gamma(0) = 0$  by definition,  $\gamma(0^+) \equiv \lim_{t \rightarrow 0^+} \gamma(t) = \tau^2$ ; this quantity is the **nugget**.
- $\lim_{t \rightarrow \infty} \gamma(t) = \tau^2 + \sigma^2$ ; this asymptotic value of the semivariogram is called the **sill**. (The sill minus the nugget,  $\sigma^2$  in this case, is called the **partial sill**.)
- Finally, the value  $t = 1/\phi$  at which  $\gamma(t)$  first reaches its ultimate level (the sill) is called the **range**,  $R \equiv 1/\phi$ .

### Examples: Spherical Variogram



b) spherical: a0 = 0.2, a1 = 1, R = 1

### Some common isotropic covariograms

Model	Covariance function, $C(t)$
Linear	$C(t)$ does not exist
Spherical	$C(t) = \begin{cases} 0 & \text{if } t \geq 1/\phi \\ \sigma^2 \left[ 1 - \frac{3}{2}\phi t + \frac{1}{2}(\phi t)^3 \right] & \text{if } 0 < t \leq 1/\phi \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Exponential	$C(t) = \begin{cases} \sigma^2 \exp(-\phi t) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Powered exponential	$C(t) = \begin{cases} \sigma^2 \exp(- \phi t ^p) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$
Matérn at $\nu = 3/2$	$C(t) = \begin{cases} \sigma^2 (1 + \phi t) \exp(-\phi t) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$

### Notes on exponential model

$$C(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t = 0 \\ \sigma^2 \exp(-\phi t) & \text{if } t > 0 \end{cases}$$

- We define the **effective range**,  $t_0$ , as the distance at which this correlation has dropped to only 0.05. Setting  $\exp(-\phi t_0)$  equal to this value we obtain  $t_0 \approx 3/\phi$ , since  $\log(0.05) \approx -3$ .
- Finally, the form of  $C(t)$  shows why the nugget  $\tau^2$  is often viewed as a **"nonspatial effect variance,"** and the partial sill ( $\sigma^2$ ) is viewed as a **"spatial effect variance."**

### The Matérn Correlation Function

- Much of statistical modelling is carried out through correlation functions rather than variograms
- The Matérn is a very versatile family:

$$C(t) = \begin{cases} \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (2\sqrt{\nu}t\phi)^\nu K_\nu(2\sqrt{\nu}t\phi) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{if } t = 0 \end{cases}$$

$K_\nu$  is the modified Bessel function of order  $\nu$  (computationally tractable)

- $\nu$  is a smoothness parameter (a *fractal*) controlling process smoothness

- How do we select a variogram? Can the data really distinguish between variograms?
- Empirical Variogram:

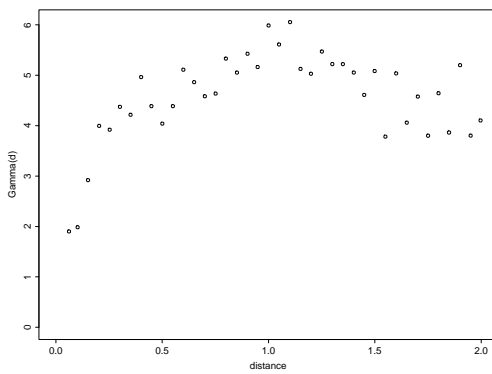
$$\gamma(t) = \frac{1}{2|N(t)|} \sum_{\mathbf{s}_i, \mathbf{s}_j \in N(t)} (y(\mathbf{s}_i) - y(\mathbf{s}_j))^2$$

where  $N(t)$  is the number of points such that  $\|\mathbf{s}_i - \mathbf{s}_j\| = t$  and  $|N(t)|$  is the number of points in  $N(t)$ .

- Grid up the  $t$  space into intervals  $I_1 = (0, t_1)$ ,  $I_2 = (t_1, t_2)$ , and so forth, up to  $I_K = (t_{K-1}, t_K)$ . Representing  $t$  values in each interval by its midpoint, we define:

$$N(t_k) = \{(\mathbf{s}_i, \mathbf{s}_j) : \|\mathbf{s}_i - \mathbf{s}_j\| \in I_k\}, k = 1, \dots, K.$$

### Empirical variogram: scallops data



### Empirical variogram: scallops data

